THE ALLELIC CORRELATION STRUCTURE OF GAINJ- AND KALAM-SPEAKING PEOPLE. I. THE ESTIMATION AND INTERPRETATION OF WRIGHT'S *F*-STATISTICS

JEFFREY C. LONG[1]

Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109

Manuscript received July 31, 1985
Revised copy accepted November 14, 1985

ABSTRACT

The internal patterning of allelic correlations in the Gainj and Kalam swidden horticulturalists of highland Papua New Guinea is examined within the context of SEWALL WRIGHT's *F*-statistic model. A multiallelic extension of the model is given first, and multivariate variance-component estimators for the parameters are suggested. Then, it is shown that the expectation of the *F*-statistic set depends on the age structure of the population and that knowledge of the population and sample age structure is critical for meaningful analysis. The array of *F*-statistics estimated jointly over five polymorphic enzyme loci reveals the following features of Gainj and Kalam population structure: (1) significant departures from panmictic expectations and (2) characteristics of a continuously distributed breeding population, rather than those expected for populations subdivided into demes with discrete boundaries. Finally, the *F*-statistics estimated for the Gainj and Kalam are briefly compared to estimates obtained from other tribal populations. It is seen that the level of differentiation observed in the Gainj and Kalam is only about one-third that observed in South American swidden horticulturalists. Consequently, some conventional wisdom regarding the interrelationship of socioecological settings and genetic structures may require reevaluation.

THE first step in the analysis of human population genetic structure often involves quantifying the departure of genotype frequencies from their panmictic expectations. WRIGHT's (1951, 1965) *F*-statistic model has been a universally popular analytical approach for this purpose because it is well suited for populations that are subdivided into smaller units, such as demes, clans or villages (*e.g.* CAVALLI-SFORZA, BARRAI and EDWARDS 1964; WORKMAN and NISWANDER 1970; WORKMAN *et al.* 1973; NEEL and WARD 1972; WIESENFELD and GAJDUSEK 1976; JORDE 1980). The model's parameters relate the departure from panmictic expectations in the total population to the Wahlund effect between subdivisions and the average departure from panmixia within subdivisions. The application presented here is designed to use observed deviations from expected genotypic proportions to identify the patterning and magnitude

[1] Present address: Human Genetics Program, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15261.

of barriers to random mating within the Gainj- and Kalam-speaking swidden horticulturalists of highland Papua New Guinea.

Application of the F-statistic model is often problematic despite its popularity. It was originally derived for codominant diallelic loci, and there has been no consensus on how it can be extended to include multiallelic loci (LI 1969; NEI 1977). In addition, the proper estimation procedures for the model's parameters have been uncertain (WEIR and COCKERHAM 1984). Difficulties arise from (1) unequal sample sizes in the subdivisions; (2) how information from multiallelic loci should be used, given that they are inadequately covered by the theory and that multivariate statistical estimation is generally more complicated; and (3) how information from different loci can be combined to increase the precision of F-statistic estimates and the power of hypothesis tests.

A review of the F-statistic model is presented before it is applied to the Gainj and Kalam. It is shown that COCKERHAM's (1969, 1973) analysis of variance (ANOVA) parameterization of the model for diallelic loci can be extended to cover multiallelic loci. The ANOVA approach to WRIGHT's theory is especially useful because it provides a clear demonstration of WRIGHT's correlational definitions of the parameters, even with multiallelic loci, and because it readily deals with problems of estimation. A method for combining F-statistic estimates from different loci is then given.

With these theoretical and methodological issues resolved, the distribution of genetic characteristics in the Gainj and Kalam is investigated. It is shown that the array of estimated F-statistics is characteristic of certain patterns of nonrandom mating and is different from either panmixia or other patterns of nonrandom mating. As a consequence, insight into the details of the population breeding structure is gained from the analysis. It is also shown that, when there is migration between the population subunits, the expectation of the F-statistic set is age structured and is quite different from the values expected with complete isolation of subunits.

THE MODEL

The F-statistic model is a hierarchical model with genes stratified at three levels: individuals (I), within subdivisions (S) and within the total population (T). It has three main parameters; F_{IT} is the correlation of uniting gametes relative to those of the total population, F_{IS} is the average over all subdivisions of the correlation of uniting gametes relative to the gametes of the subdivision and F_{ST} is the correlation of random gametes within subdivisions relative to the total population. These definitions are relative to a current population (*cf.* WRIGHT 1965, p. 401), and such correlations may arise as the consequence of a variety of evolutionary processes. The three F-statistics are interrelated as

$$(1 - F_{IT}) = (1 - F_{ST})(1 - F_{IS}). \tag{1}$$

A variety of derivations of this basic relationship are available (WRIGHT 1951, 1965; COCKERHAM 1969).

Suppose a locus has Z alleles, designated $A_1, A_2, \cdots, A_Z$, with population frequencies $p_1, p_2, \cdots, p_Z$. The variability at this locus is examined in the context of a set of random vectors, $\mathbf{x}$, each with $Z - 1$ elements. The first

TABLE 1

Scoring for a three-allele locus

Genotypes	Vectors
A_1A_1	$\begin{bmatrix}1\\0\end{bmatrix}$ $\begin{bmatrix}1\\0\end{bmatrix}$
A_1A_2	$\begin{bmatrix}1\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\end{bmatrix}$
A_1A_3	$\begin{bmatrix}1\\0\end{bmatrix}$ $\begin{bmatrix}0\\0\end{bmatrix}$
A_2A_2	$\begin{bmatrix}0\\1\end{bmatrix}$ $\begin{bmatrix}0\\1\end{bmatrix}$
A_2A_3	$\begin{bmatrix}0\\1\end{bmatrix}$ $\begin{bmatrix}0\\0\end{bmatrix}$
A_3A_3	$\begin{bmatrix}0\\0\end{bmatrix}$ $\begin{bmatrix}0\\0\end{bmatrix}$

vector element is assigned a one if the allele is type A_1, and a zero otherwise. At the second position, a one is assigned if the allele is type A_2, and a zero otherwise. And so forth until the position $Z - 1$ is reached. Thus, the vector corresponding to the allele A_Z is a vector with $Z - 1$ zeros. This approach will yield identical values for the F-statistics no matter which alleles are designated $A_1, A_2, \cdots, A_Z$. Since each person carries two alleles, two vectors are scored for each individual sampled. Although the methodology and procedures adopted in the following will remain the same for loci with any number of alleles, the subsequent derivations and explanations will be presented in terms of a three-allele locus for the purposes of brevity and concreteness. An example of the scoring procedure is given in Table 1.

The expectation of the vector $\mathbf{x}$ is $\mathbf{p} = [p_1, p_2]$ and the associated variance-covariance matrix is

$$\Sigma = \begin{bmatrix} p_1(1 - p_1) & -p_1p_2 \\ -p_1p_2 & p_2(1 - p_2) \end{bmatrix}.$$

This variance-covariance matrix can be subdivided into components corresponding to the various population subdivisions by considering the linear model

$$\mathbf{x}_{kij} = \mathbf{p} + \mathbf{a}_k + \mathbf{b}_{ki} + \mathbf{w}_{kij}, \qquad (2)$$

where $\mathbf{x}_{kij}$ indexes the jth allele, $j = 1, 2$, in the ith individual, $i = 1, \cdots, n_k$, in the kth group, $k = 1, \cdots, K$. The effects are a for groups, b for individuals and w within individuals. All effects are assumed random and uncorrelated and have the associated variance-covariance matrices $\sum_a$, $\sum_b$ and $\sum_w$. The expectation of quadratics over vectors are

$$\begin{aligned} E(\mathbf{x}_{kij}\mathbf{x}^T_{k'i'j'}) &= \mathbf{pp}^T + \Sigma && \text{if } k = k', \ i = i', \ j = j' \\ &= \mathbf{pp}^T + \Sigma_{ab} && \text{if } k = k', \ i = i', \ j \neq j' \\ &= \mathbf{pp}^T + \Sigma_a && \text{if } k = k', \ i \neq i \\ &= \mathbf{pp}^T + \Sigma_g && \text{if } k \neq k'. \end{aligned} \qquad (3)$$

For uncorrelated subdivisions, $\Sigma_g = 0$. In general,

$$\Sigma_{ab} = F_{IT} \tag{4a}$$

$$\Sigma_a = F_{ST}. \tag{4b}$$

The correlations are related to the variance-covariance component matrices (VCCM) as

$$(1 - F_{IT})\,\Sigma = \Sigma_w \tag{5a}$$

$$(F_{IT} - F_{ST})\,\Sigma = \Sigma_b \tag{5b}$$

$$F_{ST}\,\Sigma = \Sigma_a \tag{5c}$$

and

$$\Sigma = \Sigma_a + \Sigma_b + \Sigma_w. \tag{5d}$$

The F-statistics are thus

$$F_{IT} = \left(\frac{1}{2}\right) TR\left[\Sigma^{-1/2}\left(\Sigma_a + \Sigma_b\right)\Sigma^{-1/2}\right] \tag{6a}$$

$$F_{ST} = \left(\frac{1}{2}\right) TR\left[\Sigma^{-1/2}\,\Sigma_a\,\Sigma^{-1/2}\right] \tag{6b}$$

$$F_{IS} = \left(\frac{1}{2}\right) TR\left[\left(\Sigma_b + \Sigma_w\right)^{-1/2}\Sigma_b\left(\Sigma_b + \Sigma_w\right)^{-1/2}\right], \tag{6c}$$

where TR denotes the trace of a matrix. The constant preceding the right-hand side of each equation is ½ because there are two elements to the gametic vectors (*i.e.*, three alleles at the locus). Premultiplication and postmultiplication by the $-\frac{1}{2}$ power of $\sum$ and $(\sum_b + \sum_w)$ is required so that the resulting standardized VCCMs will be symmetrical.

ESTIMATION

Although the F-statistic model has been formulated for almost a half century, there has been no consensus on how to estimate its parameters. For recent reviews of the difficulties, see CURIE-COHEN (1982), ROBERTSON and HILL (1984) and WEIR and COCKERHAM (1984). The estimation procedures presented below are given in matrix notation and are obtained by the method of least squares. If we consider the single random variable associated with a diallelic locus as a one-element matrix, these formulas reduce to those presented by COCKERHAM (1969, 1973) and WEIR and COCKERHAM (1984), but they differ from the multiallelic versions given by WEIR and COCKERHAM (1984). The ease of generality to loci with any number of alleles, applicability to all three F-statistics and the availability of approximate tests for null hypotheses are advantages of the methods presented here.

Our initial objective is to estimate the matrices, $\sum$, $\sum_a$, $\sum_b$ and $\sum_w$, given in (5). This is accomplished by estimating a set of mean square and cross-product (MSCP) matrices corresponding to each level of the hierarchy.

TABLE 2

ANOVA table

Source of variation	d.f.	MSCP	Expected MSCP
Among groups	$K - 1$	MSCP (A)	$\Sigma_W + 2\Sigma_B + c\Sigma_A$
Individuals within groups	$N - K$	MSCP (B)	$\Sigma_W + 2\Sigma_B$
Within individuals	N	MSCP (W)	Σ_W

$c = 2\ (N - \frac{1}{N} \sum_k n_k^2)/(K - 1)$

$$\mathrm{MSCP}(A) = \frac{1}{K-1}\left[\sum_k \frac{\left(\sum_i \sum_j \mathbf{x}_{kij}\right)\left(\sum_i \sum_j \mathbf{x}_{kij}\right)^T}{2n_k} - \frac{1}{2N}\left(\sum_k \sum_i \sum_j \mathbf{x}_{kij}\right)\left(\sum_k \sum_i \sum_j \mathbf{x}_{kij}\right)^T\right] \tag{7a}$$

$$\mathrm{MSCP}(B) = \frac{1}{N-K}\left[\frac{1}{2}\sum_k \sum_i \left(\sum_j \mathbf{x}_{kij}\right)\left(\sum_j \mathbf{x}_{kij}\right)^T - \sum_k \frac{\left(\sum_i \sum_j \mathbf{x}_{kij}\right)\left(\sum_i \sum_j \mathbf{x}_{kij}\right)^T}{2n_k}\right] \tag{7b}$$

$$\mathrm{MSCP}(W) = \frac{1}{N}\left[\sum_k \sum_i \sum_j \mathbf{x}_{kij}\mathbf{x}_{kij}^T - \frac{1}{2}\sum_k \sum_i \left(\sum_j \mathbf{x}_{kij}\right)\left(\sum_j \mathbf{x}_{kij}\right)^T\right], \tag{7c}$$

where n_k is the number of individuals sampled in the kth subdivision and $N = \sum_k n_k$ is the total number sampled. The expectation of the MSCPs are given in Table 2, and the VCCMs are estimated by subtraction from the MSCPs. These can be denoted S_a, S_b and S_w. The matrix $\sum$ is estimated by $S = S_a + S_b + S_w$. These estimators can all be shown to be unbiased (*cf.* SCHEFFÉ 1959). The formulas given in equations 7a–c can be easily related to results from familiar ANOVA and least-squares estimation; however, they may be more easily obtained using simple gene and genotype counting methods. Alternative computational formulas are given in the APPENDIX.

MULTIPLE LOCI

The expectation of the F-statistic set is the same for all neutral loci. Thus, a battery of genetic loci can be regarded as repeated trials of the same process. This fortunate property of the model allows us to minimize the variance of F-

statistic estimates by combining information over as many independent neutral loci as have been sampled. To begin, combined VCCM are constructed in a block diagonal fashion. For example, suppose there are $l = 1, \cdots, L$ loci with corresponding VCCMs $S_{a1}, S_{a2}, \cdots, S_{aL}; S_{b1}, S_{b2}, \cdots, S_{bL}; S_{w1}, S_{w2}, \cdots, S_{wL}$. The matrices take the form

$$S_A = \begin{bmatrix} S_{a1} & 0 & \cdot & 0 \\ 0 & S_{a2} & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & S_{aL} \end{bmatrix} = \text{diag}\{S_{aL}\} \tag{8a}$$

$$S_B = \text{diag}\{S_{bL}\} \tag{8b}$$

$$S_W = \text{diag}\{S_{wL}\}. \tag{8c}$$

It is these matrices that are the subject of further analyses. Once again, the form of these matrices and the subsequent equations will remain the same, no matter how many loci are analyzed or how many polymorphic alleles are contained within each locus.

ESTIMATORS AND TEST STATISTICS

In the following, estimators for the F-statistics and test statistics, to judge statistical significance, are given. The test statistics are approximately distributed according to Wilk's Λ distribution. Formulas for this distribution are given in many textbooks (*cf.* ANDERSON 1958; MARDIA, KENT and BIBBY 1980). The approximation of the test statistics to Wilk's distribution results from the convergence of the multinomial to the multivariate normal distribution at fairly large sample sizes. In the following equations, G will denote the dimension of the matrices S_A, S_B and S_W.

F_{IS} is estimated by

$$F_{IS}^* = (1/G)TR[(S_B + S_W)^{-1/2}S_B(S_B + S_W)^{-1/2}] \tag{9a}$$

and the test statistic is

$$\Lambda_1^* = \det[\text{MSCP}(W)]/\det[\text{MSCP}(B)] \sim \Lambda(G, N - K, N). \tag{9b}$$

In this case and the following cases, the MSCP matrices will be combined matrices in an analogous fashion to the VCCM. The convention $\Lambda_i^* \sim \Lambda_{(df_1, df_2, df_3)}$ will be used to indicate that a test statistic is approximately distributed as a Wilk's random variable with df_1, df_2 and df_3 degrees of freedom.

F_{ST} is estimated by

$$F_{ST}^* = (1/G)TR[S^{-1/2}S_AS^{-1/2}] \tag{10a}$$

and the test statistic is

$$\Lambda_2^* = \det[\text{MSCB}(B)]/\det[\text{MSCP}(A)] \sim \Lambda(G, K - 1, N - K). \tag{10b}$$

Finally, F_{IT} is estimated by

$$F_{IT}^* = (1/G)TR[S^{-1/2}(S_A + S_B)S^{-1/2}], \tag{11a}$$

and the test statistic is

$$\Lambda_3^* = \det[(N - 2)\mathrm{MSCP}(B) + 2\mathrm{MSCP}(A)]/\det[(N)\mathrm{MSCP}(W)] \sim \Lambda(G, N - 1, N). \tag{11b}$$

These estimators are weighted averages of the values obtained from each of the L loci. The weight each locus receives depends solely on the number of alleles present. Ideally, equal sample sizes will be present for each locus analyzed. When this condition is not met, additional weighting schemes can be devised. One such method is given in the APPENDIX. Like all intraclass correlation coefficients, these are not necessarily unbiased. This is because they are functions of ratios of unbiased estimators, rather than unbiased estimators of ratios. Nevertheless, preliminary computer simulations (LONG 1984) have demonstrated these estimators to be unbiased over their biologically realistic range.

The variance of F-statistic estimates may be estimated via several approaches. Taylor series expansions and numerical jackknife procedures have been previously suggested in the literature (NEI and CHAKRAVARTI 1977; WEIR and COCKERHAM 1984, respectively). Taylor series approximations to the variance are recommended for single diallelic loci because analytical expressions are available and approximate confidence intervals for the F-statistic estimates can be constructed (*cf.* NETER and WASSERMAN 1974, p. 532). Extensions could be developed for multiallelic loci, but the formulas are exceedingly complicated, even for triallelic loci, and the approximations are not likely to be very accurate. Jackknife estimates of the variance are preferable in these cases. The methods for obtaining jackknife estimates are fully outlined by WEIR and COCKERHAM (1984), and the reader is referred to this paper. The variance and other statistical properties of the multivariate variance-component estimators are currently being investigated using computer simulation methods. The preliminary results indicate that the variance of F_{IT}^* is approximately $1/[(Z - 1)N]$, where N is the number of individuals sampled, and Z is the number of alleles at the locus. These results also indicate that the F-statistic estimators are not normally distributed, so that the construction of confidence intervals using the normal distribution may be misleading (LONG 1984). For this reason, simple tests of null hypotheses may be more appropriate than the construction of confidence intervals.

It is clear from WRIGHT's formulation of the F-statistic model that the parameters F_{IS} and F_{IT} are free to take either positive or negative values depending on whether there is a deficit or excess of heterozygotes. For this reason, two-sided hypothesis tests are recommended. It is also clear from WRIGHT's work that the parameter F_{ST} is necessarily positive, and this leads us to recommend a one-sided hypothesis. *It should be noted that* $\mathrm{F}_{\mathrm{ST}}^*$ *is a random variable and it may take negative values, unlike the parameter* F_{ST}. Such negative estimates can be interpreted as F_{ST} = zero, because it is impossible to reject the one-sided null hypothesis when F_{ST}^* is less than or equal to zero.

APPLICATION

The Gainj and Kalam speakers are two neighboring groups of tribal horticulturalists living on the northern fringe of Papua New Guinea's central highlands (WOOD, SMOUSE and LONG 1985; LONG 1984; J. C. LONG *et al.*, 1986). These two contiguous populations are linguistically distinct but ecologically and culturally very similar. They are both subdivided into a series of local groups ranging in size from about 20 to 200 individuals. Within each local group, members share access to the territory's garden land and forest resources. In keeping with long-standing anthropological usage in New Guinea, these groups will be referred to as parishes (HOGBIN and WEDGWOOD 1952).

Gainj and Kalam parishes constitute the most easily identified units of residential structure. The F-statistic analysis presented here is therefore structured relative to the parish level of subdivision. Nevertheless, it is uncertain to what extent parish membership represents a barrier to random genetic exchange. There is a very high rate of marital exchange between parishes, and all of the parishes included in this analysis are in very close geographic proximity (WOOD, SMOUSE and LONG 1985). The major questions to be answered are (1) whether or not there is significant genetic differentiation between parishes, and should there be differentiation; and (2) what are the barriers to random genetic exchange, both within and among parishes? Thus, for the purposes of this study, F_{IS} is the correlation of alleles within individuals, relative to the allelic array of the parish; F_{ST} is the correlation of random alleles within parishes, relative to the entire population; and F_{IT} is the correlation of alleles within individuals, relative to the allelic array of the entire population.

Five loci, red cell acid phosphatase (*ACP*), glutamic-pyruvate transaminase (*GPT*), adenosine deaminase (*ADA*), haptoglobin (*Hp*) and Group Specific Component (*Gc*) were analyzed. The first four of these loci each have two alleles, whereas *Gc* is triallelic. Of the 52 genetic systems for which these people have been characterized (LONG *et al.*, 1986), these five loci were the only ones that met the following criteria for selection. First, they are codominant systems. Second, there are typings in all 21 parishes for which blood samples have been collected. Third, in the total population, the average frequency of the most common allele is <0.90. The first condition ensures that all single-locus genotypes are observable, and the second provides an adequate number of degrees of freedom for hypothesis testing. The third condition is important for a variety of more subtle reasons reviewed below.

Computer simulation studies have shown the variance of the F-statistic estimates to be high and unstable with respect to allele frequency when the most common allele is above the 0.90 frequency level. By contrast, when the most frequent allele is below this level, the variance of the F-statistic estimates are related to sample sizes, but nearly independent of allele frequencies (LONG 1984). It is also the case that F-statistic estimates can be substantially biased when the level of polymorphism is low and small samples have been drawn. Finally, F-statistic estimates from loci with low levels of polymorphism are not strictly comparable with estimates from loci with high polymorphism for purely mechanical reasons. F_{IS} is defined only in the interval $[-q/(1 - q); 1.0]$, where q is the frequency of the least frequent allele (CURIE-COHEN 1982). Thus, if q is the frequency of a rare variant, the negative portion of F_{IS}'s range will be far more restricted than if q is the frequency of a common morph. Both

TABLE 3

Estimated *F*-statistics

Locus	F_{IS}	F_{ST}	F_{IT}	*N*
ACP	0.0177	0.0144*	0.0318	570
ADA	0.0836**	0.0310*	0.1120**	570
GPT	0.0439	0.0426**	0.0847	569
Hp	−0.0313	0.0227**	−0.0079	470
Gc	0.0474	0.0109**	0.0544	522
Total	0.0364**	0.0224**	0.0568**	

* $P < 0.05$; ** $P < 0.01$.

theoretical (COCKERHAM 1973) and empirical studies (SMOUSE, NEEL and LIU 1983) have indicated that F_{IS} values as low as −0.10 are reasonable for human populations. Consequently, it will not be instructive to examine loci for which F_{IS} values are excluded from the region because of low polymorphism.

RESULTS

F-statistics estimated individually for the five loci, and combined for the set, are presented in Table 3. Sample sizes and significance levels (*P* values) are also given. The significance levels should be interpreted cautiously, because the distributional requirements are achieved only at large sample sizes. Nevertheless, it is comforting that applications of various approximate χ^2 tests for the individual loci (SNEDECOR and IRWIN 1933; LI and HORVITZ 1953; COCKERHAM 1969) yield nearly identical results. The significance levels for the multiple locus estimates must be accepted even more tentatively, because there is a considerable amount of missing data at the haptoglobin locus. The primary reason for this is that ahaptoglobinemia reaches a fairly high frequency in this population (approximately 20%). As a conservative precaution, the haptoglobin sample size was used to determine the number of degrees of freedom for hypothesis tests. Despite these inadequacies of the testing procedure, there are few cases of ambiguous significance, so the interpretation of the results is not impaired.

The single locus F_{IS} coefficients range from −0.0313 (*Hp*) to 0.0836 (*ADA*). The single locus F_{ST} coefficients are far more homogeneous, ranging from 0.0109 (*Gc*) to 0.0426 (*GPT*). There is nothing revealed by the interlocus patterns of either F_{IS} or F_{ST} that would suggest that selective forces are operating on these loci. However, five loci is far too few a number to expect such patterns to emerge.

The single-locus estimates of F_{IT} span −0.0079 (*Hp*) to 0.1120 (*ADA*). The negative F_{IT} value observed at the *Hp* locus is not statistically significant, and only the extremely high observation at *ADA* reaches unequivocal significance. Finally, the *Hp* locus illustrates an important point. The positive and significant value of F_{ST} is countered by a negative value of F_{IS}. Thus, the genotypes in the entire population are nearly in panmictic proportions, despite the fact that nonrandom associations of alleles exist within and among parishes.

The multiple-locus estimates of all three F-statistics are seen to be positive and significantly different from zero. We see that F_{IT} is elevated by both within and among subdivision deviations from panmixia, and roughly to the same extent. The implication of the values for the Gainj and Kalam is complicated by the fact that there are many evolutionary pressures affecting genotype proportions simultaneously (*cf.* WORKMAN 1969). Nevertheless, a careful examination of the pattern of allelic correlations does shed light on the Gainj and Kalam breeding structure.

INTERPRETATION OF F_{ST}

Despite the close geographical proximity and high marital exchange rate between parishes, the overall estimate of F_{ST} (0.0224; $P < 0.01$) indicates that significant subpopulation structure exists within the study group. This value is in excess of the value ($F_{ST} = 0.001$) inferred from a heterogeneity χ^2 analysis of blood-group phenotypes observed in 15 clans of New Guinea highlanders at Bundi (MALCOLM, BOOTH and CAVALLI-SFORZA 1971), and it is reasonably close to the level ($F_{ST} = 0.0201$) observed in 19 "villages" of the North and South Fore linguistic groups in the eastern highland (WIESENFELD and GAJDUSEK 1976). The low level of clan differentiation is consistent with theoretical expectations (MORTON, IMAIZUMI and HARRIS 1971), and the higher level of differentiation of "villages" within different linguistic groups is not unexpected, given that linguistic barriers are usually barriers to mate exchange as well.

Beyond these very generalized statements, it is difficult to interpret F_{ST} values. GAJDUSEK and his colleagues (GAJDUSEK and ALPERS 1972; SIMMONS *et al.* 1972; WIESENFELD and GAJDUSEK 1976) have pointed out that we do not know the precise relationship between the existing social units and the population breeding structure. They note that before European contact, true tribal organization or formalized tribal leadership did not exist anywhere in the New Guinea highlands. Traditionally, the people resided in small hamlets ranging in size from a few to several families. The people were so geographically restricted that few people even knew the extent of their own language group. Larger political units arose from loose affiliations of hamlets resulting from marriage ties, property rights, adoption of children and frequent exchange of members. These affiliations were both temporally and spatially plastic, and they even extended across linguistic boundaries. GAJDUSEK and ALPERS (1972, p. 7) conclude, "if there is anything resembling tribal organization today, it follows from government determination of linguistic boundaries, appointment of village leaders, (as luluai and tultul), conduct of census, and European construct of a named tribe."

Both the precontact population pattern and the European overlay are apparent in the Gainj and Kalam today. The parishes that we identify correspond closely to administrative units within the Gainj census division. The Gainj census division is, itself, a conglomeration of people; most are Gainj speakers, but many Kalam and some Maring speakers are included as well. The larger parishes we have sampled are actually collections of named, but smaller, places. These smaller places correspond closely to the hamlets described by GAJDUSEK and ALPERS (1972). Our demographic information reveals that these smaller places may realign with different parishes episodically. The question thus arises, does our value of F_{ST} reflect variation accrued by discrete breeding units, or rather, does it reflect heterogeneity within a continuous breeding population (*cf.* WRIGHT 1943) that has been arbitrarily subdivided for administrative purposes? Because F_{ST} is a simple summary measure that can arise in either of these situations, it cannot answer this question (WRIGHT 1965). Nevertheless, the other F-statistics offer much to resolve this issue.

INTERPRETATION OF F_{IS}

The value of F_{IS} (0.0364; $P < 0.01$) estimated here is enigmatic at first glance. Virtually all empirical studies and theoretical considerations lead us to expect negative F_{IS} values in tribal populations (NEEL and WARD 1972; COCKERHAM 1973; WORKMAN *et al.* 1973; JORDE 1980; SMOUSE, NEEL and LIU 1983). The factors that affect allelic correlations include finite size, subdivision, differential fertility and sex differences in gamete frequencies, assortative mating, migration and mixtures of subdivisions. It will be shown, however, that the value of F_{IS} observed in the Gainj and Kalam is not unreasonable, given our sampling framework and the juxtaposition of those forces that would arise from the precontact breeding structure.

Consider a population with discrete generations that is subdivided into completely isolated units. Allowing for separate sexes and random mating within subdivisions, it is well known (COCKERHAM 1973) that, after one generation, F_{IS} is constant across generations (although F_{ST} and F_{IT} can continuously increase) and F_{IS} is negative. NEI (1973) has extended this model to allow for differential fertility of males and females (including concomitant sex differences in allele frequencies). He finds the expectation of F_{IS} to be $E(F_{IS}) = -[1/(8n_m) + 1/(8n_f)]$, where n_m and n_f are, respectively, the effective numbers of males and females. When n_f and n_m are equal, $E(F_{IS})$ is $-1/(2n_e)$, but when they differ, F_{IS} becomes more negative. This formulation still assumes random mating with respect to consanguinity, which includes brother-sister and other types of relative matings. Any avoidance of consanguinity serves to make F_{IS} even more negative.

Turning to the Gainj and Kalam, the average effective population size of the parishes is estimated to be roughly 31 individuals. This figure is simply one-third of the average parish size. In addition to this small number, male and female contributions to the breeding population are uneven, due to polygyny, and marriages between close relatives are consciously avoided (JOHNSON 1982). Given these considerations, we would expect F_{IS} to be on the order of -0.02 if the above population structure model were realistic.

The most obvious departures of the Gainj and Kalam from this simple model are age structuring and migration between subdivisions. The effects of these factors are modeled by dividing the population into parental (postmigratory) and offspring (premigratory) age categories and finding the expectation of F_{IS} within each group. We shall assume that, within each parish, the frequency of some arbitrary allele $f(A_1)$, is p_m in male parents and p_f in female parents and that genotypes in both sexes are close to their respective panmictic proportions. The parental group genotypes (combined males and females) will not be in panmictic proportions but rather,

$$f(A_1A_1) = p^2 + \sigma^2_{p(m,f)}$$
$$f(A_1A_2) = 2[p(1-p) - \sigma^2_{p(m,f)}]$$
$$f(A_2A_2) = (1-p)^2 + \sigma^2_{p(m,f)},$$

where p is the average allele frequency, and $\sigma^2_{p(m,f)}$ is the between-sex variance in allele frequency. This variance in allele frequency can be manipulated to the form $\sigma^2_{p(m,f)} = (1/4)(p_m - p_f)^2$ (*cf.* LI 1976, p. 64). Thus, the parents have a Wahlund sort of effect reducing the number of heterozygotes, and this effect can easily be expressed in terms of the sex difference in allele frequency. The expectation of F_{IS} in parents is then $E_p(F_{IS}) = \sigma^2_{p(m,f)}/p(1-p)$.

The offspring distribution of genotypes is well known (ROBERTSON 1965);

$$f(A_1A_1) = p_mp_f = p^2 - \sigma^2_{p(m,f)}$$
$$f(A_1A_2) = p_m(1-p_f) + p_f(1-p_m) = 2[p(1-p) + \sigma^2_{p(m,f)}]$$
$$f(A_2A_2) = (1-p_m)(1-p_f) = (1-p)^2 - \sigma^2_{p(m,f)},$$

where $p_m = p - (1/2)(p_f - p_m)$ and $p_f = p + (1/2)(p_f - p_m)$. The expectation of F_{IS} in the offspring portion of the population is then $E_o(F_{IS}) = -\sigma^2_{p(m,f)}/p(1-p)$, exactly the same magnitude as, but opposite in sign to, the parental value.

When the population subdivisions are not completely isolated, the expectation of F_{IS} is age structured. The expectation of F_{IS} is then

$$E(F_{IS}) = \int_0^{\omega} c(x)F(x)dx \simeq \sum_0^{\omega} c_xF_x, \tag{12}$$

where ω is the oldest attainable age, $F(x)$ and F_x are the respective continuous and discrete age-specific F_{IS} functions and $c(x)$ and c_x are, respectively, the continuous and discrete probability density functions giving the relative proportion of the population in each age category (CAVALLI-SFORZA and BODMER 1971, p. 298). It is apparent that estimation of F_{IS} requires knowledge of the population age structure and that the sampling framework must reflect this knowledge in order to avoid seriously biased results. These are important considerations in the study population because there is a high interparish migration rate ($m = 0.45$) and, by and large, postmarital residence is patrilocal, so migrant individuals are most often female. This system almost guarantees there will be a sex difference in allele frequencies within the adult members of a parish.

To help resolve the unusually high value of F_{IS} observed in the Gainj and Kalam, the sample has been divided into two categories. For convenience, these are designated postmigratory and premigratory, although membership is determined solely on the basis of age. The premigratory category includes indi-

TABLE 4

Age-structured *F*-statistics

Age category	F_{IS}	F_{ST}	F_{IT}	N
Premigratory	0.0255	0.0392**	0.0637**	113
Postmigratory	0.0395**	0.0263**	0.0770**	320

N = sample size of Hp locus.
** $P < 0.01$.

viduals between the ages of 10 and 24, whereas the postmigratory category includes those between 25 and 64. This age division is pragmatic. Many people have already migrated by age 24 (Wood 1980), but this lumping is necessary to obtain adequate sample sizes. As a consequence, we shall underestimate the age effect.

The results of the analysis are presented in Table 4. F_{IS} and F_{ST} are, respectively, 0.0395 and 0.0263 in the postmigrants, whereas they are 0.0255 and 0.0392 in the premigrants. As expected, F_{IS} is elevated in the postmigrants, and F_{ST} is elevated in the premigrants. We expect a generational difference in F_{ST}, and we note that the difference between the postmigrant and premigrant samples is 0.0129. This difference would be expected (assuming a generational difference of $1/2n_e$) if the average parish effective size were about 39 individuals. This expected value of n_e is somewhat greater than our rough estimate from the parish sizes ($n_e \simeq 31$). It should be noted, however, our figure for the generational difference in F_{ST} is probably an underestimate due to the lumping of age classes, and most likely, neither method for estimating n_e is very efficient.

Finally, we note that the relationship between the sampling units and the population breeding structure will affect the expectation of F_{IS}. Consider first the possibility that the parishes represent discrete demes. The expectation of F_{IS} will reflect the pattern of mate exchanges within and between demes, and it will be given by (11). Negative values of F_{IS} will be expected (at least in the younger age categories), unless there is positive assortative mating (Cockerham 1973). Now consider the alternative that the parishes are merely administrative units and that the larger ones are conglomerates of smaller demes. Equation (12) is no longer sufficient to predict F_{IS}, because it is inflated in all age categories by the Wahlund effect of lumping smaller demes. The first possibility is unlikely. There is a deficit of heterozygotes (*i.e.* positive F_{IS}) in all age categories, and consanguinous marriages are actively avoided. The alternative possibility is consistent both with the data and the probable precontact breeding structure. It is additionally supported by cultural data and some further analysis of the genetic markers.

If the marriage system is truly exogamous with respect to the breeding units, but the larger parishes are, in fact, conglomerates of smaller breeding units, then some marriages that are actually exogamous with respect to the true breeding units will appear endogamous with respect to the administrative units (parishes). This proportion will be expected to increase with parish size, an observation already reported by JOHNSON (1982).

We would also expect F_{IS} to be lower (possibly negative) in smaller parishes and greater (positive) in larger parishes. To test this theory, the 21 parishes were divided into two groups. The first group contained 11 parishes with census sizes of less than 99 individuals, and the second group was composed of the 10 parishes with census sizes of more than 99 people. F_{IS} was then calculated separately for each set of parishes. The respective values were −0.0148 and 0.0458, once again agreeing with the predictions.

The positive value of F_{IS} is therefore consistent with the precontact breeding structure sampled within the framework of the postcontact administrative units. It is not consistent with the notion that the parishes are discrete breeding units.

INTERPRETATION OF F_{IT}

We can conclude from the observed value of F_{IT} (0.0568; $P < 0.01$) that there is a moderately high correlation between alleles within individuals relative to the allelic array of the total population. This correlation is specific to the age categories appropriate to our sample and does not represent the level of correlation existing in the population as a whole. This is because F_{IT} and F_{ST}, like F_{IS}, have age-structured expectations.

The effect of age structure on F_{ST} has already been worked out in detail for the migration matrix model of population structure (ROGERS and HARPENDING 1983). The general implications of age structure for F_{ST} are heuristically reiterated below. Consider a subdivided population with migration between the subunits. A cohort of individuals is born within each subdivision; they grow to maturity, and migration between the subdivisions takes place. Mating follows migration, and the cycle is complete with the production of a new cohort within each subdivision. An ontogeny of genetic variation parallels this process. There is a certain level of differentiation among subdivisions at the beginning of the cycle. This level of differentiation is diminished after migration. A new increment is then accrued with the formation of the next generation. The magnitude of this sampling effect is approximately $1/(2n_e)$. Thus, genetic differentiation between the subdivisions oscillates with respect to the life cycle. When the homogenizing effects of migration and the diversifying effects of sampling are of equal magnitude, the level of differentiation remains constant across generations with respect to the phase of the life cycle. This leads to two conclusions: (1) a generational difference in F_{ST} will be expected, regardless of whether or not a between-generation equilibrium has been attained; and (2) in age-structured populations, the population value of F_{ST} will be an average over age-specific expectations of F_{ST} and the population age distribution. Finally, given the effect of age structure on both F_{IS} and F_{ST}, the age structuring of F_{IT} is obvious.

DISCUSSION

Despite the popularity of WRIGHT's (1951, 1965) F-statistic model, numerous alternative models and a plethora of estimation procedures have been suggested in recent years (*cf.* SPIELMAN, NEEL and LI 1977). This is because the model parameters, as defined by WRIGHT, are difficult to estimate and because there has not been an obvious extension of the model to multiallelic loci. The extension presented here is appealing because of its straightforward connection with the theory for diallelic loci and because of its simple preservation of WRIGHT's correlational definitions of the parameters. As with COCKERHAM's original (1969, 1973) use of the ANOVA format, approximate hypothesis testing procedures are available with this multiallelic extension. In addition, the treatment of information from a sample of different genetic loci follows without difficulty.

Application to genotypic data collected on Gainj- and Kalam-speaking people of Papua New Guinea has elucidated the relationship between their current social units and their population breeding structure. The pattern of the F-statistics indicates that the basic postcontact social units do not display the genetic characteristics of discrete demes, but rather, the larger parishes appear to be conglomerates of smaller breeding units. The current genetic structure thus reflects the traditional social organization in highland New Guinea described by GAJDUSEK and ALPERS (1972). That is a population fragmented into numerous small hamlets with extremely ephemeral social cohesion.

It is well known that estimates of F_{ST} are sensitive to the way the subdivisions are defined (NEEL and WARD 1972; JORDE 1980), and it is apparent that estimates of F_{IS} are as sensitive. The consequence of lumping smaller breeding units is then the inflation of F_{IS} and the deflation of F_{ST}. It is interesting that careful analysis of the F_{IS} component has yielded the most insight into the Gainj and Kalam population structure. In retrospect, this should not be too surprising, as COCKERHAM (1973) has noted, F_{IS} attains equilibrium rapidly and is more reflective of the immediate population breeding structure than the other parameters of the model.

The observation of positive F_{IS} in the Gainj and Kalam distinguishes them from all other anthropological populations yet studied. For example, in the Yanomama and Makiritare Indians of South America and the Papago and Zuni tribes of North America, negative values of F_{IS} are observed (NEEL and WARD 1972; WORKMAN *et al.* 1973, 1974). Negative values of F_{IS} are characteristic of populations subdivided into discrete breeding units, and consequently, these groups are characterized by different breeding structures. F_{ST} has been estimated for many more populations than F_{IS} (see JORDE 1980), and it has been claimed that high values of this statistic ($F_{ST} > 0.04$) are found primarily among tropical agriculturalists of restricted mobility, with substantially lower values characterizing hunter-gatherers of greater mobility (HARPENDING 1974). Although the Gainj and Kalam estimate of F_{ST}, 0.0224, is higher than that for the !Kung hunter-gatherers, 0.007, studied by HARPENDING and JENKINS (1973, 1974), it is only about one-third that observed for the horticultural Yanomama Amerindians, 0.063 (SPIELMAN, NEEL and LI 1977). This is so, despite the fact that the Gainj and Kalam are far more sedentary than the Yanomama. It is reasonable to hypothesize that this lower value of F_{ST} derives from the continuous breeding structure of the Gainj and Kalam, rather than from the discrete Yanomama subdivisions.

Consequently, we find a more intricate relationship between socioecological settings and population breeding structures than has heretofore been realized. This conclusion prevents us from making sweeping generalizations about human evolution and variation from the study of a few "typical" populations, but it presents the opportunity to pose more detailed questions about the covariation of human biological and social variation. From an optimistic viewpoint, we can expect to find new insight into the evolutionary significance of the forces of the population breeding structure by examining and comparing different groups.

I should like to thank GEORGE ESTRABROOK, C. C. LI, FRANK LIVINGSTONE, JAMES V. NEEL, PAULA NEMETH and PETER SMOUSE for their comments on earlier drafts of this paper. Useful suggestions were also provided by two anonymous reviewers. The fieldwork depended on many Gainj and Kalam people, the Papua New Guinea Institute of Medical Research and PETER SMOUSE, PATRICIA JOHNSON, and JAMES WOOD. Financial support of this research was provided by National Institutes of Health grants NIH R01 GM30135 and NIH T32 GM07544 and a Dissertation grant from the Horace H. Rackham School of Graduate Studies, University of Michigan.

LITERATURE CITED

ANDERSON, T. W., 1958 *An Introduction to Multivariate Statistics.* John Wiley & Sons, New York.

CAVALLI-SFORZA, L. L., I. BARRAI and A. W. F. EDWARDS, 1964 Analysis of human evolution under random genetic drift. Cold Spring Harbor Symp. Quant. Biol. **29:** 29.

CAVALLI-SFORZA, L. L. and W. F. BODMER, 1971 *The Genetics of Human Populations.* W. H. Freeman, San Francisco.

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C. 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

CURIE-COHEN, M., 1982 Estimates of inbreeding in a natural population. Genetics **100:** 339–38.

GAJDUSEK, D. C. and M. ALPERS, 1972 Genetic studies in relation to kuru. Am. J. Hum. Genet. **24** (Suppl): s1–s38.

HARPENDING, H. C., 1974 Genetic structure of small populations Annu. Rev. Anthropol. **3:** 229–243.

HARPENDING, H. C. and T. JENKINS, 1973 Genetic distances among Southern African populations. In: *Methods and Theories of Anthropological Genetics,* Edited by M. H. CRAWFORD and P. L. WORKMAN. University of New Mexico Press, Albuquerque.

HARPENDING, H. C. and T. JENKINS, 1974 !Kung population structure. In: *Genetic Distance,* Edited by J. F. CROW and C. F. DENNISTON. Plenum Press, New York.

HOGBIN, I. and C. WEDGWOOD, 1952 Local grouping in Melanesia. Oceania **23:** 241–276.

JOHNSON, P. L., 1982 Gainj kinship and social organization. Ph.D. Dissertation, University of Michigan, Ann Arbor.

JORDE, L. B., 1980 The genetic structure of subdivided human populations: review. In: *Current Developments in Anthropological Genetics, Vol. 1. Theory and Methods,* Edited by H. Mielke and M. H. Crawford. Plenum Publishing Corp., New York.

LI, C. C., 1969 Population subdivision with respect to multiple alleles. Ann. Hum. Genet. **33:** 23–29.

LI, C. C., 1976 *Population Genetics,* Boxwood Press, Pacific Grove, California.

LI, C. C. and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

LONG, J. C., 1984 The estimation of genetic variation and divergence: application to Gainj and Kalam speakers of highland New Guinea. Ph.D. Dissertation, University of Michigan, Ann Arbor.

LONG, J. C., J. M. NAIDU, H. W. MOHRENWEISER, H. GERSHOWITZ, P. JOHNSON, J. W. WOOD and P. E. SMOUSE, 1985 Genetic characterization of Gainj and Kalam speaking people of Papua, New Guinea. Am. J. Phys. Anthrop. In press.

MALCOLM, L. A., P. B. BOOTH and L. L. CAVALLI-SFORZA, 1971 Intermarriage patterns and blood group gene frequencies of the Bundi people of the New Guinea highlands. Hum. Biol. **43:** 187–199.

MARDIA, K. V., J. T. KENT and J. M. BIBBY, 1980 *Multivariate Analysis.* Academic Press, New York.

MORTON, N. E., Y. IMAIZUMI and D. E. HARRIS, 1971 Clans as genetic barriers. Am. Anthropol. **73:** 1005–1010.

NEEL, J. V. and R. H. WARD, 1972 The genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by *F*-statistics (including a comparison with the Makiritare and Xavante). Genetics **72:** 639–666.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3323.

NEI, M., 1977 *F*-statistics and the analysis of gene diversity in subdivided populations. Ann. Hum. Genet. **41:** 22–233.

NEI, M. and A. CHAKRAVARTI, 1977 Drift variances of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. Theor. Pop. Biol. **3:** 460–465.

NETER, J. and W. WASSERMAN, 1977 *Applied Linear Statistical Models.* Richard D. Irwin, Inc., Homewood, Illinois.

ROBERTSON, A., 1965 The interpretation of genotypic ratios in domestic animal populations. Anim. Prod. **7:** 319–324.

ROBERTSON, A. and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

ROGERS, A. R. and H. C. HARPENDING, 1983 Population structure and quantitative characters. Genetics **105:** 985–1002.

SCHEFFÉ, H., 1959 *The Analysis of Variance.* John Wiley & Sons, Canada.

SIMMONS, R. T., J. J. GRAYDON, D. C. GAJDUSEK, M. P. ALPERS and R. W. HORNABROOK, 1972 Genetic studies in relation to kuru. II. Blood-group genetic patterns in Kuru patients and populations of the eastern highlands of New Guinea. Am. J. Hum. Genet. **24** (Suppl): s39–s71.

SMOUSE, P. E., J. V. NEEL and W. LIU, 1983 Multiple-locus departures from panmictic equilibrium within and between village gene pools of Amerindian tribes at different stages of agglomeration. Genetics **104:** 133–153.

SNEDECOR, G. and M. R. IRWIN, 1933 On the chi-square test for homogeneity. Iowa State J. Sci. **8:** 7–81.

SPIELMAN, R. S., J. V. NEEL and F. H. F. LI, 1977 Inbreeding estimation from population data: models, procedures and implications. Genetics **8:** 3–371.

WEIR, B. S. and C. C. COCKERHAM, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

WEISENFELD, S. L. and D. C. GAJDUSEK, 1976 Genetic structure and heterozygosity in the kuru region, eastern highlands of New Guinea. Am. J. Phys. Anthropol. **4:** 177–190.

WOOD, J. W., 1980 Mechanisms of demographic equilibrium in a small human population, the Gainj of Papua New Guinea. Ph.D. Dissertation, University of Michigan, Ann Arbor.

WOOD, J. W., P. E. SMOUSE and J. C. LONG, 1985 Sex-specific dispersal patterns in two human populations of highland New Guinea. Am. Nat. **125:** 747–768.

WORKMAN, P. L., 1969 The analysis of simple genetic polymorphisms. Hum. Biol. **41:** 97–114.

WORKMAN, P. L., H. HARPENDING, J. M. LALOUEL, C. LYNCH, J. D. NISWANDER and R. SINGLETON, 1973 Population studies on southwestern Indian tribes. VI. Papago population structure: a comparison of genetic and migration analyses. In: *Genetic Structure of Populations,* Edited by N. MORTON. University of Hawaii Press, Honolulu.

WORKMAN, P. L. and J. D. NISWANDER, 1970 Population studies on southwestern Indian tribes. Am. J. Hum. Genet. **22:** 24–49.

WORKMAN, P. L., J. D. NISWANDER, K. S. BROWN and W. C. LEYSHON, 1974 Population studies on Southwestern Indian tribes. IV. The Zuni. Am. J. Phys. Anthropol. **41:** 119–132.

WRIGHT, S., 1943 Isolation by distance. Genetics **28:** 114–138.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen (Lond.) **1:** 323–34.

WRIGHT, S., 1965 The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution **19:** 39–420.

Communicating editor: B. S. WEIR

APPENDIX: COMPUTATIONAL METHODS

Formulas for the mean square and cross-product and VCCMs are developed below from simple gene and genotype counting procedures. Once again, the methods will be presented for a three-allele locus, although the expansion to accomodate loci with any number of alleles is straightforward. To begin, suppose there are Z alleles at a locus, $A_1, A_2, \cdots, A_Z$, with the usual $\frac{1}{2}[Z(Z+1)]$ corresponding genotypes. Let n_{kij} be the number of individuals sampled in the kth subdivision with the genotype A_iA_j, and let n_k be the total number of individuals sampled in the kth subdivision. Thus,

$$n_k = \sum_i n_{kii} + \sum_{i<j} n_{kij}$$

The number of individuals sampled from the total population with the genotype A_iA_j will be denoted

$$N_{ij} = \sum_i n_{kij}$$

and the number of individuals sampled in the total population will be

$$N = \sum_i N_{ii} + \sum_{i<j} N_{ij}.$$

Now, let $n_{ki\cdot}$ be the number of A_i alleles sampled in the kth subdivision

$$n_{ki\cdot} = 2n_{kii} + \sum_{j:j\neq i} n_{kij}$$

and, accordingly, let $N_{i\cdot}$ be the number of A_i alleles sampled in the total population.

$$N_{i\cdot} = \sum n_{ki\cdot}$$

The estimators for the mean-square and cross-product matrices are then

$$\mathrm{MSCP}(W) = 1/(2N) \begin{bmatrix} N_{12} + N_{13}; & -N_{12} \\ -N_{12}; & N_{13} + N_{23} \end{bmatrix}$$

$$\mathrm{MSCP}(B) = 1/(2[N - K]) \left[\begin{bmatrix} 4N_{11} + N_{12} + N_{13}; & N_{12} \\ N_{12}; & 4N_{22} + N_{12} + N_{23} \end{bmatrix} - \sum_k (1/n_k) \begin{bmatrix} n_{k1\cdot}^2; & n_{k1\cdot}n_{k2\cdot} \\ n_{k1\cdot}n_{k2\cdot}; & n_{k2\cdot}^2 \end{bmatrix} \right]$$

$$\mathrm{MSCP}(A) = 1/(2[K - 1]) \left[\sum_k (1/n_k) \begin{bmatrix} n_{k1\cdot}^2; & n_{k1\cdot}n_{k2\cdot} \\ n_{k1\cdot}n_{k2\cdot}; & n_{k2\cdot}^2 \end{bmatrix} - 1/N \begin{bmatrix} N_{1\cdot}^2; & N_{1\cdot}N_{2\cdot} \\ N_{1\cdot}N_{2\cdot}; & N_{2\cdot}^2 \end{bmatrix} \right]$$

The VCCMs are estimated as

$$S_W = \mathrm{MSCP}(W)$$

$$S_B = (1/2)[\mathrm{MSCP}(B) - \mathrm{MSCP}(W)]$$

$$S_A = (1/C)[\mathrm{MSCP}(A) - \mathrm{MSCP}(B)],$$

where C is given in Table 2, and $S = S_A + S_B + S_W$.

The $-(1/2)$ power of a full rank matrix, say S, can be found from the relationship

$$E^{-1}SE = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdot & 0 \\ 0 & \lambda_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \lambda_n \end{bmatrix} = \mathrm{diag}\{\lambda_i\},$$

where E is a square matrix whose columns are the eigenvectors of S, and Λ is a diagonal matrix whose elements are the corresponding eigenvalues of S. $S^{-1/2}$ is then

$$S^{-1/2} = E\Lambda^{-1/2}E^{-1} = E\ \mathrm{diag}\{1/\sqrt{\lambda_i}\}E^{-1}$$

The F-statistics are then solved for as in (9–11).

The matrix methods for estimating the F-statistics jointly over a set of $l = 1, \cdots, L$ independent loci serve to illustrate the connection between the test statistics and Wilk's distribution, but the joint estimates can be equivalently obtained as simple weighted averages of the individual values from the L loci. Let G_l be the number of elements in the gametic random vector from the lth locus; thus $G = \sum_l G_l$. Now allow F_l to be the estimate of F_{IS}, F_{ST} or F_{IT} from that locus. It is not important to distinguish among the F-statistics, because the weighting scheme is the same for all three measures. The weighted estimate is then

$$F = \left[\sum_l G_l F_l \right] \Big/ G.$$

Supposing the L loci are represented by slightly different sample sizes, N_l, the weighting scheme can be expanded as follows:

$$F = \left[\sum_l N_l G_l F_l \right] \Big/ \left[\sum_l N_l G_l \right].$$

This weighting scheme is based on the rationale that the variance of F_{IT} is approximately $1/[(Z - 1)N]$; when all N_l are equal, it reduces to the standard form.